Br. J. clin. Pharmac. (1982), **14**, 325–331

STATISTICAL SIGNIFICANCE TESTS

D.R. COX
Department of Mathematics, Imperial College, London SW7 2BZ

Introduction

Statistical methods for the analysis of data can be classed as descriptive or as probabilistic. The former methods include tabulation and graphical display, of great importance both in detailed analysis and also in presentation of conclusions. Probability enters statistical analysis in various ways. The most obvious is that one recognizes explicitly that conclusions drawn from limited data are uncertain and that therefore it is a good thing to measure and control probabilities of drawing the wrong conclusions, partly, but by no means solely, as a precaution against overinterpretation. The statistical significance test is the most widely known device concerned with such probabilities and is the subject of the present paper.

The nature, use and misuse of significance tests have been much discussed in both statistical and nonstatistical circles. The view taken in this paper is that such tests play an important but nevertheless strictly limited role in the critical analysis of data.

There is an extensive mathematical literature on the theory of such tests and on the distributions appropriate for special tests. The mathematics can largely be separated from critical discussion of the underlying concepts, that being the concern of the present paper.

Nature of statistical significance tests

The first element in a significance test is the formulation of a hypothesis to be tested, often called the null hypothesis. It may help to consider first how to proceed when random variation can be ignored. We might, for instance, be interested in a hypothesis about how some biological mechanism 'really works'. We would do the following:

(i) Find one (or more) test observations (or sets of observations) whose values can be predicted from the hypothesis under examination. For instance the hypothesis might predict that a certain set of relationships between 'response' and log dose are in fact parallel straight lines. The test observations would then be such as to allow assessments of nonlinearity and nonparallelism.

(ii) Make the relevant observations.

(iii) If the test observations do not agree with the predictions, the hypothesis is rejected, or at least modified. If the test observations agree with the prediction, we conclude that the data are, in this respect, consistent with the hypothesis. Note, though, that we cannot conclude that the hypothesis is true, because there may be other hypotheses that would have led to the same test data.

While the general idea of significance tests follows essentially the same steps (i–iii) there are, however, two important differences. First the observations are subject to non-trivial variability, natural random variability and/or measurement 'errors' of various kinds. Secondly the hypotheses are typically not scientific (biological) hypotheses, but rather hypotheses about the values of parameters describing 'true' values applying to populations of measurements. Of course, in some fields at least, these parameters will be linked to scientific models or interpretations of the system under study.

We now adapt the procedure (i)–(iii) as follows.

(i) Find a test statistic whose probability distribution when the null hypothesis is true is known, at least approximately. Also the test statistic is to be such that large, or possibly extreme (large or small), values of the statistic point to relevant departures from the null hypothesis.

(ii) Calculate the value of the test statistic for the available data.

(iii) If the test statistic is in the extreme of its probability distribution under the null hypothesis, there is evidence that the null hypothesis is untrue. If the test statistic is in the centre of its distribution, the data are consistent with the null hypothesis.

More quantitatively, we calculate from the distribution of the test statistic, the probability P that a deviation would arise by chance as or more extreme than that actually observed, the null hypothesis being true. The value of P is called the (achieved) significance level. If $P < 0.05$, we say the departure is significant at the 0.05 level, if $P < 0.01$ that the departure is significant at the 0.01 level, and so on. In applications it is nearly always enough to find P fairly roughly.

A natural and fundamental question now arises. Why consider at all deviations more extreme than that observed? Why assess a hypothesis partly on the basis of what has *not* been observed.' One answer is as follows. Suppose that we were to accept the observations as just decisive evidence against the null hypothesis. Then we would be bound to accept more

0306–5251/82/090325–07 $01.00

extreme data as evidence against that hypothesis. Therefore P has a direct, though hypothetical, physical interpretation. It is the probability that the null hypothesis would be falsely rejected when true, were the data just decisive against the hypothesis. Note particularly that the interpretation is hypothetical. In the first place, at least, we use P as a measure of evidence, not as a basis for immediate tangible action (see also Appendix E).

A simple example

Consider the following much simplified example. To compare a drug D with a placebo C, 100 patients each receive both drugs, in random order and with appropriate design precautions. For each patient an assessment is made as to which is the more effective treatment, ties not being allowed. It is observed that for 60 patients D was better and for 40 patients C was better.

Now the null hypothesis would typically be that the long run proportion of preferences for D is ½. Other null hypotheses could, however, arise; for instance if a large independent study under comparable conditions had shown that 80% preferred D, another interesting null hypothesis would be that the long run proportion in the present study is 0.8.

The test statistic is the observed number of preferences for D. Under the 50% null hypothesis the probability distribution of the number of preferences for D can be found; independence of the individuals is assumed. The distribution, which is roughly a normal distribution of mean 50 and standard deviation 5, is that for the number of heads observed in 100 tosses of a 'fair' coin. The distribution could also be found by computer simulation although, except possibly for didactic reasons, this is a poor idea when a mathematical treatment is available. In complicated problems however, where a mathematical treatment is not available, it is a good idea to simulate data from a 'null' system close to that under study. Conformity of the 'real' with the 'simulated' data can be examined at least informally.

Now the probability P of a deviation as or more extreme than that observed is, at first sight, the probability of observing 60, or 61, or . . . , or 100 and is about 0.029.

But there is a difficulty here; what if we observed not 60 preferences for D and 40 for C but 40 for D and 60 for C? If under those circumstances we would still have examined consistency with the null hypothesis, our significance level P must take account of extreme fluctuations in both directions, the test becomes a so-called two-tailed or two-sided test and in this instance P becomes $2 \times 0.029 \simeq 0.057$.

That is we have

$$P_{\text{one-sided}} = \text{prob}\{60, \text{or } 61, \text{or } \ldots, \text{or } 100\} \simeq 0.029,$$
$$P_{\text{two-sided}} = \text{prob}\{60, \text{or } 61, \text{or } \ldots, \text{or } 100; 0, \text{or } \ldots, \text{or } 39, \text{or } 40\} \simeq 0.057$$

There has been some controversy as to which is appropriate when *a priori* we expect any difference to favour D and this expectation is confirmed in the data. To an appreciable extent the matter is one of convention. In reading the literature, check to see which has been used. In the present paper we use two-sided tests and leave the matter with the following brief comments.

(a) In some problems the two-sided P is clearly appropriate, because neither direction of effect is biologically impossible.

(b) It will be rather rare that a substantial observed effect in an unexpected direction can be dismissed without discussion. Further, retrospective explanations of unexpected effects can often be put forward. These are the arguments for regarding the two-sided test as the one for routine use.

(c) It is largely conventional whether one says that the two-sided P is say 0.1 and the effect is in the expected (or unexpected) direction, or whether one quotes the one-sided P of 0.05.

Section 3 describes the nature of null hypotheses, and we then go on to discuss various points of interpretation.

The nature of null hypotheses

Clearly for significance tests to be of much use, sensible and interesting null hypotheses should be tested. A full classification of null hypotheses will not be attempted, but it is useful to have the following types in mind:

(i) null hypotheses of a technical statistical character to the effect that such and such special assumptions are satisfied, such as that distributions have a particular mathematical form or that preliminary transformation of data is unnecessary. We do not consider these further here;

(ii) null hypotheses closely related to some scientific hypothesis about underlying biology, such as, in genetics, that two genes are on different chromosomes. These are of critical importance in certain kinds of research, but are perhaps of less direct importance in clinical trials, although in complex trials with elaborate data, it may be feasible to formulate such hypotheses;

(iii) null hypotheses of homogeneity. These typically are that a certain effect is the same, for example in several independent studies or in different subgroups of patients. An emphasis in statistical work is frequently on designing and analysing individual investigations so that these investiga-

tions, in isolation, lead to convincing conclusions. Yet, clearly, judgment about any issue should depend on all available information; the most convincing conclusions are those backed by several independent studies and preferably by different kinds of evidence. Tests of homogeneity play some role here. It is a reasonable comment on some statistical analyses that not enough emphasis is placed on the combination of evidence from several sources.

(iv) Finally, and perhaps most commonly, there are hypotheses that assert the equality of effect of two or more alternative treatments, for example, a drug and a placebo. One way of looking at such hypotheses is that they divide the possible true states of affairs into two (or more) qualitatively different regimes. So long as the data are consistent with the equality of two treatments, the data do not firmly establish the sign of any difference between them. From this point of view significance tests address the question: how firmly do the data establish the direction of such and such an effect?

Some comments on interpretation

There follow a number of miscellaneous comments on the interpretation of significance tests.

(i) There are clear distinctions between statistical significance and the more important notions of biological significance and practical significance. Statistical significance is concerned with whether, for instance, the direction of such and such an effect is reasonably firmly established by the data under analysis. Biological significance is concerned with underlying scientific interpretation in terms of more fundamental mechanisms. Practical significance is concerned with whether an effect is big enough to affect practical action, e.g. in the treatment of patients. The three kinds of significance are linked but in a relatively subtle way.

(ii) Failure to achieve an interesting level of statistical significance, i.e. the observation of values reasonably consistent with the null hypothesis, does not mean that practically important differences are absent. Thus in the Example of Section 2 the observation of 55 preferences for D and 45 for C would lead to $P_{\text{two-sided}} = 0.37$ and to the conclusion that the data are just such as to be expected if the null hypothesis were true. On the other hand it can be shown that long run effects ranging from 45% to 65% preferences for D are consistent with the data (at the 5% significance level) and, depending entirely on the context, these might correspond to differences of practical importance. By constrast with 1000 patients a split 550 for D and 450 for C would be a highly significant departure from the null hypothesis ($P_{\text{two-sided}} \simeq 0.0018$), the range of preference for D consistent with the data is from 52% to 58% and yet it might well be that these would all be of little practical importance. In this instance the direction of the effect is firmly established but its magnitude is such as to make the effect of little practical importance; such instances of observational overkill are probably rare in clinical trials.

(iii) The strong implication of (ii) is that, for important effects at least, it is necessary to calculate so-called confidence limits for the magnitude of effects and not just values of P. This is of crucial importance. It is very bad practice to summarise an important investigation solely by a value of P.

(iv) Failure to achieve an interesting level of significance in a study does not mean that the topic should be abandoned. Significance tests are not intended to inhibit the free judgment of investigators. Rather they may on the one hand warn that the data alone do not establish an effect, and hence guard against overinterpretation and unwarranted claims, and on the other hand show that an effect is reasonably firmly proved.

(v) If an effect is established as clearly corresponding to a bigger difference than can be accounted for by chance, two conditions have to be satisfied before the difference can be given its face-value interpretation as, say, an improvement due to the new drug under test. One condition is that the calculation of statistical significance is technically sound in that no unreasonable detailed assumptions have been made in computing the probabilities. The most common fault is, in effect, to overlook an important source of variability and thereby to underestimate the uncertainty in the conclusions. It is a key role of careful design to ensure that, so far as is feasible, all major sources of uncertainty are controlled and that their magnitude can be assessed. Specialised statistical advice may be needed on this point. This is related to the second and more serious possibility that the effect is a bias or systematic error and not a genuine 'treatment' effect. Biases are relatively much more likely in observational studies, and comparisons involving historical rather than concurrent controls, than in carefully controlled experiments involving randomisation and, so far as feasible, firmly enforced 'blindness'. To a very limited extent biases can be investigated and corrected by more elaborate statistical analysis (see Appendix D).

(vi) Pooling of information from independent studies is very important, subject to tests of homogeneity. If there is clear evidence of heterogeneity between studies this needs to be explained in subject-matter terms if at all possible.

(vii) If a number of independent studies are done, possibly in different centres, and only those yielding an effect significant at say the 0.05 level are reported, it is clear that a distorted picture may emerge. The criterion for publication should be the achievement of

reasonable precision and not whether a significant effect has been found; it may be, however, that investigations yielding interesting new conclusions merit publication in greater detail than investigations whose outcome is predominantly negative or confirmatory, but that is a separate issue.

The choice of tests and levels: multiple testing: choice of sample size

A question frequently asked is 'what significance level should be used, 0.05, 0.01 or what?' The view taken in the previous discussion is that the question does not arise, at least in the sense that we may use the observed value of P as a measure of the conformity of the data with the null hypothesis. What overall interpretation is made or what practical action is taken depends not only on the significance level P, but also on other information that may be available, including biological plausibility on the magnitude of effect present and, in a decision-taking context, on the consequences of various modes of action.

Nevertheless some rough guide for qualitative interpretation is useful and the following seems fairly widely accepted, sensible, and leads to some standardization of procedure:

$P \geq \frac{1}{10}$: accord with the null hypothesis is as good as could reasonably be expected;

$P \leq \frac{1}{100}$: there is, for most purposes, overwhelming evidence against the null hypothesis;

intermediate: there is some evidence against the null hypothesis.

Values of P should be quoted approximately and rigid borderlines avoided. Thus in terms of the interpretation of data it would be absurd to draw a rigid borderline between $P = 0.051$, not significant at the 0.05 level, and $P = 0.049$, significant at the 0.05 level, even though in decision-making contexts with strictly limited data such borderlines do have to be adopted.

Some special problems arise when several statistical tests are applied to the same data. We can distinguish a number of situations:

(i) Scientifically quite separate questions can be addressed from the same data. This raises no special problems. Note, however, that if one of the analyses led to doubts about data quality, the other analyses would become suspect too.

(ii) The data may be searched by examining different kinds of effect and ultimately focusing attention on the most significant. For instance one might classify the patients in various ways (age, sex, etc.) to look for differential effects. Or we may, entirely legitimately, examine several measures of response: death rate (all causes), death rate (particular causes), 'minor' critical event rate, and so on. Quite clearly if many different effects are examined it is quite likely that a superficially quite significant result will emerge. Individual P values will apply to each effect in isolation, but to judge whether there is overall evidence of an effect, a so-called allowance for selection is essential in judging significance. A rough rule is to multiply the smallest P value by the number of effects examined before selecting the most significant. For instance suppose that an overall analysis of a clinical trial comparing drug D with placebo C shows little evidence of a difference. Analyses are made for men and women, for 'young' and 'old' patients and for two different Centres, say taking the 8 combinations of sex-age-Centre. The older women at Centre 2 show a difference significant at about the 0.05 level. But this is the largest out of 8 comparisons and therefore, after the rough correction for selection, is to be assigned a value of P of about $8 \times 0.05 = 0.4$, i.e. it is just such as would be expected by chance. Note that if the investigation had been set up specifically to investigate older women in Centre 2, $P = 0.05$; the magnitude of the effect is the same in both context. This general point is especially important because major trials are complex with many measures of response, categories of patient etc.

(iii) The third possibility is that essentially the same scientific issue is tackled by trying a number of tests similar in spirit but different in detail. Thus to compare two samples one might consider the t test, the sign test, the Wilcoxon test, the Fisher and Yates score test, and so on. Each is appropriate under slightly different conditions and usually, except possibly for the sign test, the results will be very similar; the differences arise primarily from the relative weights attached to extreme observations. This blunderbuss procedure of trying many tests is wrong if what is done is concealed. In principle, for each specific scientific question there should be a single test in the light of the error structure present. If a search through a number of test statistics is carried out, an allowance for selection is essential. The simple correction outlined above is much too crude, being really appropriate for statistically independent tests rather than for closely related tests. If the procedure were used, which is not recommended, computer simulation to attain the appropriate correction for selection seems the most practicable procedure.

(iv) A rather more specialised issue concerns the normal use of significance tests in trials in which the possibility of stopping the trial before its nominal completion date is determined at least in part by formal data-dependent rules. We regard this topic,

in its practical and theoretical aspects, as outside the present discussion.

(v) A simpler issue is the use of hypothetical significance test calculations to determine an appropriate scale of investigation, e.g. the number of patients or study duration. The argument is to require that if a difference of an amount judged important is present, then the probability often denoted by β should be at least say 0.9 that a difference significant at the 0.05 level often denoted by α is achieved; the values of β and α can of course be altered. Often an estimate has to be made of such matters as the overall critical event rate likely to be encountered. In many ways a simpler approach is to work with the standard error of the estimated difference likely to be achieved, aiming to control this a suitable level. Some calculation of this sort is highly desirable to prevent the setting up of investigations either absurdly small or on an absurdly extravagant scale. Several somewhat arbitrary elements are involved in the above calculations: in principle the scale of investigation could be settled by decision analysis, costing the effort spent in the study and the consequences of wrong decisions. This is but part of the broader issue of the role of formal decision analysis in medical work, and a brief comment on this follows.

Relation with decision analysis

An accompanying paper in this series (Spicer, 1982) deals with the interesting topic of decision analysis in a medical context. Superficially there is a connexion between the so-called two decision problem, i.e. deciding between the two alternative courses of action, and the significance test of an appropriate null hypothesis. In fact, at least from the viewpoint of the present paper, the procedures are best regarded as quite distinct. The significance test is concerned with some aspects of assessing information provided by an investigation. A decision or diagnostic rule indicates an appropriate action, based on a synthesis of the information provided by data, of so-called prior information available from other sources and of utilities measuring the consequences of various actions in given circumstances. The decision rule requires a greater variety of quantitative input than does a significance test and in a sense reaches a correspondingly stronger conclusion.

Reference

SPICER, C.L. (1982). Statistical decision theory and clinical trials. *Br. J. clin. Pharmac.*, **14** (in press).

Appendix

Some further issues, including some idealised misconceptions

There follow a few statements involving significance tests, together with brief comments. Probably none of A–C has ever been made in quite the form given here, but the statements may nevertheless pinpoint potential errors of interpretation. D and E outline some further issues of importance, marginal to the main theme of the paper. By 'effect' is meant, for example, a difference in survival rate as between two alternative regimes. The null hypothesis is that this difference is zero.

A. *The effect is significant at the 1% level: that is, the probability of the null hypotheses of zero effect is only 1%.* Not so! This is not what significant at the 1% level means (see Section 2). So-called Bayesian significance tests do allow the calculation of probabilities of hypotheses being true (or false), but require a great deal of extra information. Most of the general points in the paper apply also to Bayesian significance tests; see Appendix E.

B. *Of two independent studies of the same topic, one gives no significant effect and the other gives significance at the 1% level. Thus the two studies are inconsistent.* Not necessarily so! A modest effect and a rather larger effect in the same direction will often satisfy the conditions stated. The magnitudes of the effects should be examined, and their mutual consistency considered. This is usually done by testing the null hypothesis of homogeneity, that the magnitudes are equal in the two studies. Subject to consistency, a pooled estimate of the effect can be calculated.

C. *An effect is highly significant (say at 1% level) but is unexpected and apparently inexplicable on biological grounds. There is thus a conflict between statistics and biology; biology should win.* Up to a point, this is reasonable enough. The conflict is not, however, between statistics and biology, but between the specific data and the biology: the statistical analysis serves to make the conflict more explicit and rather less a matter of personal judgment than might otherwise be the case. Resolution of the conflict will often require new data but sometimes it may be sought on the initial data via one or more of the following: (i) there is a bias (systematic error) in the data; (ii) the calculation of P is in error, e.g. by overlooking some source of variability; (iii) an extreme chance fluctuation has arisen; (iv) the initial biological understanding is faulty. Possibilities (i) and (ii) can be the subject of further statistical investigation. Possibility (iii) is unverifiable. As to (iv), the

traffic between biological understanding and more empirical studies should be two-way; to disregard all empirical studies in conflict with prior expectation would be very dangerous.

D. *Biases can be more important than random errors and biases are not susceptible to statistical analysis.* Biases are indeed likely to be relatively important in large studies, where the effect of random variation is controlled by the averaging of responses over a large number of patients. It is one of the primary objectives of careful design to eliminate biases, by balancing or by randomizing (with concealment) all important sources of variability. In observational studies, for instance when historical controls are used instead of the much more desirable randomized concurrent controls, biases are appreciably more likely and certainly the main sources of potential bias should be examined and if necessary adjustments calculated. In broad terms, if the source of bias can be identified and is short of a total confounding with the treatment effect under study, then there is a possibility of largely removing the bias by suitable statistical analysis. On general grounds it is, of course, very undesirable that interpretation should depend in an essential way on elaborate analysis: direct appreciation of the meaning and limitations of the data is thereby inhibited.

Biases are of three broad types: (i) those arising from systematic differences between the two (or more) arms of the trial, involving patient characteristics on entry; (ii) those concerned with the procedures for measurement of response and other variables; (iii) those concerned with the implementation of the treatment and with the care of the patient during any subsequent period of observation.

Randomization is intended to ensure the absence of the first kind of bias. Sometimes, however, randomization may be defective; further, in observational studies appreciable differences may arise between a new 'treated' group and a historical control group and these will need adjustment by some process of standardization, for example by analysis of covariance. Significance tests are now more complex but not different in principle. A major concern is that once it is clear that two groups are unbalanced with respect to an observed property (e.g. age or sex) the possibility of a difference with respect to an unmeasured though important property is enhanced. This would lead to a bias not removable by analysis.

Biases connected with measurement can arise either when 'blindness' is impossible or seriously ineffective or when different observers, laboratories, etc. are involved. It is difficult to suggest an effective practical method of dealing with the first kind of bias: independent confirmation of particularly sensitive observations may sometimes be feasible, possibly on a sample of patients. Interobserver and inter-laboratory systematic differences can in principle be estimated and eliminated, provided of course the rather disastrous situation is avoided wherein, say biochemical measurements in the two treatment arms are made in two different laboratories. External evidence on the degree of standardization achieved will always be valuable.

Biases of the third kind, concerned with treatment implementation, are in some ways the most awkward. A conventional answer is that the usual comparison in a randomized clinical trial tests the total consequences of being assigned to one treatment arm rather than to the other. Thus subsidiary effects, such as differences in general level of care, different levels of compliance, unacceptability of side effects, etc. are part of the consequences of the treatments under study. From this point of view, biases are spirited away by definition. It is clear that this approach may sometimes be quite unsatisfactory, both scientifically and for decision making about the treatment of individual patients. If the conditions of this experiment are quite different from those obtaining in practical use, what is sometimes called an external systematic error arises. Practical ways of dealing with these matters of 'treatment bias' are not well developed. In principle, if properties associated with treatment administration and subsequent care can be measured, they may be either deliberately varied by design or alternatively measured for each patient or for small groups of patients and their effect on the major response assessed by suitable analysis. The difficulties of measuring accurately such aspects as compliance are well known.

E. *The Bayesian approach to statistical tests allows the incorporation of external evidence, that is always available, and avoids the conceptual difficulties of ordinary significance tests.* It would be out of place in the present paper to write at length on the foundations of statistical inference. These foundations have for many years been under active discussion and debate: such discussion should be taken as evidence of the intellectual vigour of the subject and not at all as implying an unsoundness in application.

Briefly, significance tests as described in the main part of the paper aim to summarise what the data tell about consistency with a hypothesis or about the direction of an effect. The P-value has, before action or overall conclusion can be reached, to be combined with any external evidence available and, in the case of decision-making, with assessments of the consequences of various actions. Clearly the P-value is a limited aspect: note also point (iii) of some comments on interpretation where the importance of calculating limits of error for the magnitude of an effect is stressed.

It appears superficially very attractive to strengthen interpretation and to bring in the external

information, by use of the so-called Bayesian approach. In my opinion, there is room for different approaches in different contexts, but the arguments are fairly persuasive against the Bayesian tests in most applications to the analysis and interpretation of data. First it may often help to keep separate the impact of data and external information. Secondly, to put the external information in the required quantitative form raises formidable conceptual and practical difficulties. (In some cases, such as certain problems of diagnosis, external information is in the form of a statistical frequency, and then there is no dispute that the Bayesian approach should be applied.)

Finally in problems in which the direction of an effect is under study and the external information is relatively weak, the one-sided P value is approximately the (Bayesian) probability that the time effect has the opposite sign to the effect in the data and the two approaches are largely reconciled.